BEBR
FACULTY WORKING
PAPER NO. 1512

M-Estimation of Multivariate Regressions

Roger Koenker
Stephen Portnoy

THE LIBRARY OF THE
NOV 28 1988
UNIVERSITY OF ILLINOIS

College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois, Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 1512

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

October 1988

M-Estimation of Multivariate Regressions

Roger Koenker, Professor
Department of Economics

Stephen Portnoy, Professor
Department of Statistics

M-Estimation of Multivariate Regressions

Roger Koenker[1] and Stephen Portnoy[2]

Departments of Economics and Statistics
University of Illinois
Champaign, IL 61820

October 1988

Abstract

Robust alternatives to the seemingly unrelated regression (SUR) estimator of Zellner (1962) are proposed for the classical multivariate regression model. These weighted M-estimators achieve an asymptotic covariance matrix analogous to that of the SUR estimator. Comparisons are made with the efficient estimator in the case of elliptically contoured distributions. The l_1, least absolute deviation, case is discussed in detail, and some remarks on estimation of the weighting matrices are offered. In contrast to recent work of Hampel, *et al* (1986), Rousseeuw(1987) and Oja(1983), the methods studied below are *not* affine equivariant; some remarks on the potential significance of this failing conclude the paper.

1. Introduction

Consider the classical multivariate regression model

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix} \tag{1.1}$$

with m equations and n observations on each equation, which we will express more succinctly as

$$y = X\beta + u \ .$$

When u is multivariate Gaussian $G(0, \Omega \otimes I)$ and β is an unknown $p = \sum_{i=1}^{m} p_i$ vector, it is well known that the ordinary least squares estimator

[1]This research was partially supported by NSF grant SES-8707169.
[2]This research was partially supported by NSF grant DMS-8802555.

$\hat{\beta} = (X'X)^{-1}X'y$ is inefficient relative to the (Gauss-Markov) generalized least squares estimator $\tilde{\beta} = (X'(\Omega^{-1} \otimes I)X)^{-1} X'(\Omega^{-1} \otimes I)\, y$. The former has covariance matrix

$$\hat{V} = V(\hat{\beta}) = (X'X)^{-1}X'(\Omega \otimes I)\, X(X'X)^{-1}$$

while the latter boasts,

$$\tilde{V} = V(\tilde{\beta}) = (X'(\Omega^{-1} \otimes I)X)^{-1}\,.$$

The difference $\hat{V} - \tilde{V}$ is positive semi-definite. Zellner (1962, 1963) contains the seminal analysis of this situation.

Similarly, it is easy to show under analogous conditions that the ordinary least absolute deviation (l_1) estimator, $\hat{\beta}$, which minimizes

$$R(b) = \sum_{i=1}^{m} \sum_{j=1}^{n} |y_{ij} - x_{ij}b_i|$$

has asymptotic covariance matrix of the form $V(\hat{\beta})$, but with Ω replaced by

$$\Omega = (\omega_{ij}) = \frac{E\ \text{sgn}\ (u_{ik})\ sgn\ (u_{jl})}{4f_i(0)f_j(0)}\,.$$

The numerator of ω_{ij} may be regarded as an l_1 correlation based on orthant probabilities between the errors in the i^{th} and j^{th} equation and the terms in the denominator are the marginal densities of these error terms evaluated at their medians. Since the latter are inversely proportional to the scale of the marginal distributions, Ω may be regarded as an l_1-covariance matrix. The bivariate version of the numerator has been considered in Blomqvist (1950); see also Devlin, *et. al.* (1975).

In light of the least squares results it is natural to ask: can we construct a generalized l_1 estimator which has an asymptotic covariance matrix of the form $V(\tilde{\beta})$? In the next section we investigate a rather broad class of weighted M-estimators which achieve a generalized version of this objective, and we shall see that a particular weighted l_1-estimator is an important special case. Since these estimators use one-dimensional kernels, Section 3 investigates their efficiency compared to the fully multivariate asymptotically optimal estimators. We consider elliptically contoured error distributions and specialize specifically

to multivariate t-distributions. The basic conclusions are that although the one-dimensional methods can have arbitrarily small efficiency, this tends to occur only when the asymptotic variance tends to zero. Thus, the simple one-dimensional methods (particularly, the appropriate l_1-estimator) will generally achieve quite reasonable asymptotic performance. Some practical comments on estimating the asymptotic covariance matrix for the l_1-estimator are made in Section 4, and a discussion of the lack of full affine equivariance is presented in Section 5.

2. M-Estimation of Multivariate Regression

Slight departures from Gaussian behavior of u can, of course, produce arbitrarily large disturbances in the behavior of the least-squares estimators referred to in the previous section. To achieve some degree of robustness against such departures from normality we might consider estimators which minimize

$$R_0(b) = \sum_{i=1}^{m} \sum_{j=1}^{n} \rho(y_{ij} - x_{ij} b_i) \ . \tag{2.2}$$

The ordinary l_1 estimator is an important special case. Estimation of the m-variate location and scatter model is also an important special case where $X_i \equiv 1_n$, an n-vector of ones and β is a m-vector of location parameters. Under mild conditions on ρ, minimizing $R_0(b)$ is equivalent to solving the equations

$$\sum_{j=1}^{n} \psi(y_{ij} - x_{ij} b_i) x_{ij} = 0 \qquad i = 1, \cdots, m \tag{2.3}$$

for $\psi = \rho'$. We will refer to estimators which utilize such one-dimensional kernels as ordinary M-estimators; in the location-scatter problem the terminology "coordinatewise M-estimator" might be used. Like the ordinary least-squares estimator they can be computed one equation (coordinate) at a time.

It should also be remarked at this stage that most of the attractive choices for ρ involve some scale estimation to achieve scale invariance. For example,

for the leading case of the Huber M-estimator,

$$\rho(z) = \begin{cases} \frac{1}{2}z^2 & |z| \leq k \\ k|z| - \frac{1}{2}k^2 & |z| > k \end{cases}$$

we require some (scale-equivariant) scale estimators s_i: $i = 1, \cdots, m$, e.g. the median absolute deviation from the l_1-fit, which can be used to rescale the objective function. In these cases we should presume that

$$\rho(y_{ij} - x_{ij}b_i) = \rho_0((y_{ij} - x_{ij}b_i)/s_i)$$

for some standardized ρ_0 and the rescaling by s_i is implicitly subsumed into the function ρ defined above. Of course, in the case of the l_1 estimator, scale invariance requires no preliminary estimation of scale. We consider this case in more detail in the Section 4.

To relax the implausible and potentially dangerous Gaussian hypothesis on u in Section 1 we will assume:

CONDITION A1. *The m-vectors $u_j = (u_{1j}, u_{2j}, \cdots, u_{mj})'$ are independent and identically distributed with joint distribution function F.*

Further conditions on F are required for particular choices of ρ and ψ, but since we are not particularly concerned here with minimal conditions on F, independent of the choice of ρ and ψ, we will simply require:

CONDITION A2. *The matrix*

$$R \otimes I = (E\psi(u_{ik})\psi(u_{jl})) = (\rho_{ij}\delta_{kl})$$

is positive definite, and the marginal distributions $F_i(u_i)$: $i = 1, \cdots, m$ of F, satisfy

$$\phi_i = \int_{-\infty}^{\infty} \psi'(u)dF_i(u)$$

for constants $0 < \phi_i < \infty$, $i = 1, \cdots, m$, and $E\psi(u_{ik}) = 0$ for $i = 1, \cdots, m$, $j = 1, \cdots, n$.

Note that in the least-squares case $\rho(u) = \frac{1}{2}u^2$, so R is simply the usual covariance matrix of the u_i, while $\phi_i \equiv 1$. In the l_1 case, R is the "orthant

probabilities correlation matrix" of covariances of the signs of the errors, while $\phi_i = 2f_i(0)$.

The asymptotic theory of the ordinary M estimator is immediately obtained from the asymptotically linear representation of the M-estimator for each equation,

$$\hat{\beta}_i - \beta_i = n^{-1}(\phi_i D_{ii})^{-1} X_i' \psi_i + o_p(n^{-1/2}) \qquad i = 1, \cdots, m \qquad (2.1)$$

where $D_{ii} = \lim n^{-1} X_i' X_i$ and $\psi_i = (\psi(u_{ij}))$, $i = 1, \cdots, m$. The joint asymptotic normality of these vectors follows immediately as in single equation context. A typical block of the covariance matrix is

$$\text{Cov}((\hat{\beta}_i - \beta_i), (\hat{\beta}_j - \beta_j)) = n^{-2}\phi_i^{-1}\phi_j^{-1}\rho_{ij}D_{ii}^{-1}X_i'X_jD_{jj}^{-1} + o_p(n^{-1})$$

Thus, the covariance matrix for the entire vector $(\hat{\beta} - \beta) = ((\hat{\beta}_i - \beta_i))$ may be written as

$$\hat{V} = (X'X)^{-1}X'(\mathbf{\Sigma} \otimes I)\, X(X'X)^{-1}$$

where $\mathbf{\Sigma} = \Phi^{-1} R \Phi^{-1}$ with $\Phi = \text{diag}(\phi_i)$. It might be noted that we can also write,

$$\hat{V} = (X'\Psi X)^{-1}X'(R \otimes I)X(X'\Psi X)^{-1}$$

where $\Psi = \Phi \otimes I$. Clearly the block diagonality of X as well as the Kronecker product form of Ψ is essential to the "simplification" above. The latter form for the asymptotic covariance matrix of the l_1 estimator has recently been derived by Kuester (1987).

As we observed above, it is natural to ask whether we can improve upon the asymptotic performance of this ordinary M-estimator, designing a generalized M-estimator which would achieve asymptotic covariance matrix,

$$\tilde{V} = (X'(\mathbf{\Sigma}^{-1} \otimes I)X)^{-1}.$$

This objective is easily achieved if we simply replace the "normal equations" of the unweighted objective function, which we may express in more compact form as,

$$X'\psi(b) = 0$$

with the *weighted* normal equations

$$X'\Psi(R^{-1} \otimes I)\psi(b) = 0. \tag{2.2}$$

A natural question at this point is whether or not there is an optimization problem which implies (2.2), but differentiating (2.2) with respect to b and noting that the resulting matrix is *not* symmetric, resolves the question negatively. Solving equation (2.2) poses no serious computational burdens beyond the obvious burden of dimensionality. Our main result is the following asymptotic representation of $\tilde{\beta}_n$, the estimator solving 2.2.

THEOREM 2.1. *In the multivariate linear model (1.1), suppose A1 and A2 hold. Then*

$$\tilde{\beta}_n - \beta = (X'\Psi(R^{-1} \otimes I)\Psi X)^{-1} X'\Psi(R^{-1} \otimes I)\psi(0) + o_p(n^{-1/2}) . \tag{2.3}$$

where $\psi(0) = (\psi(u_{ij}))$.

Proof. Consider the normalized gradient,

$$g(\delta) = n^{-1/2} X'\Psi(R^{-1} \otimes I)\psi(\delta)$$

where $\psi(\delta) = (\psi(u_{ij} + x_{ij}\delta_i/\sqrt{n}))$, an mn-vector. Familiar arguments from Ruppert and Carroll (1980) and Bickel (1975) imply for fixed $L > 0$,

$$\sup_{\|\delta\| < L} \|g(\delta) - g(0) - E(g(\delta) - g(0))\| = o_p(1) . \tag{2.4}$$

Further, $\tilde{\delta} = \sqrt{n}(\tilde{\beta} - \beta) = O_p(1)$, $Eg(0) = 0$, and $g(\tilde{\delta}) = o_p(1)$. Finally expanding $\psi(\cdot)$ we have

$$\sup_{\|\delta\| < L} \|Eg(\delta) - n^{-1}X'\Psi(R^{-1} \otimes I)\Psi X\,\delta\| = o_p(1) \tag{2.5}$$

so substituting $\tilde{\delta}$ in (2.5) and then in (2.4) completes the argument. □

An immediate application of this result is the asymptotic normality of $n^{1/2}(\tilde{\beta} - \beta)$, which has mean zero (since $E\psi(0) = 0$). The asymptotic covariance matrix of $(\tilde{\beta} - \beta)$ is

$$\tilde{V} = (X'\Psi(R^{-1} \otimes I)\Psi X)^{-1} = (X'(\Sigma^{-1} \otimes I)X)^{-1}$$

Note that each component $(\tilde{\beta}_i - \beta_i)$ is expressed in Theorem 2.1 as a weighted sum of n independent components. Our design assumption insures that these summands satisfy the Lindeberg condition, cf. Koenker and Bassett (1978).

As in the classical least-squares case it is important to consider the consequences of replacing Ψ and R in (2.2) by estimates. However, similar arguments to those in the classical context yield an identical asymptotic theory provided $\hat{\Sigma} \to \Sigma$ in probability. In subsequent work we hope to explore the practical consequences of various estimation schemes for Σ.

3. Comparisons with Optimal Estimators in the Elliptically Contoured Case

While solving (2.2) provides an asymptotic improvement over the naive M-estimator, this method still depends on a one-dimensional kernel. Since the problem is inherently multidimensional, this poses the question of how much one is sacrificing for the sake of simplicity. Two comments can be made here.

First, the results of Portnoy ((1977) and, especially, (1979), section 1) suggest that if there is only small dependence between the equations, a one-dimensional kernel with a small amount of redescent provides the first order correction to the optimal estimator. Thus, there is little sacrifice of efficiency if the dependence is small. If the dependence is large, however, improvements can be made by using fully multivariate estimators; for example, the maximum likelihood estimator for model (1.1). Comparisons are somewhat difficult to make in the completely general case, but the elliptically contoured case provides relatively clear and simple comparisons. Consider $u = (u_1, \cdots, u_n)$ as a matrix of a sample of size n from a multivariate density, f, on $\mathbf{R}^m$ which is elliptically contoured with parameter Λ. That is, Λ^{-1} is the "precision matrix" which can be used to transform the distribution of u_j to be spherically symmetric. The matrix Λ is not uniquely defined, but is only determined up to a positive multiplicative constant. Thus, when variances exist, we will generally specify the constant by taking

$$\Lambda = \mathrm{Cov}(u_j) \quad . \tag{3.1}$$

Clearly, the results do not depend on having a finite variance, but the specification in (3.1) will permit direct comparisons to be made. The specific

example considered below will take u_j to have a multivariate t-distribution (with covariance Λ), and will emphasize the case where $m = 2$. The first result provides the optimal asymptotic covariance matrix as the inverse Fisher Information.

THEOREM 3.1. *Consider the elliptically contoured case above. Define the function g on $\mathbf{R}^+$ by*

$$g(u'\Lambda^{-1}u) = -\log f(u) \quad \text{for} \quad u \in \mathbf{R}^m \quad . \tag{3.2}$$

Assume appropriate regularity conditions for the maximum likelihood estimator to have an (optimal) asymptotic covariance matrix equal to the inverse of the Fisher Information Matrix. (For example, general conditions applicable to this SUR problem can be found in Theorem 4.2 (p. 194) of Ibragimov and Has'minskii (1981)). Then this optimal covariance matrix is

$$V^* = c^* (X'(\Lambda^{-1} \otimes I) X)^{-1} \tag{3.3}$$

where $\frac{1}{c^*} = \frac{4}{m} E \, \|u_j\|^2 (g'(\|u_j\|^2))^2$.

Proof. First consider the spherically symmetric case ($\Lambda = I$). Using the coordinate notation of Section 1, the log-likelihood can be written

$$-L(\beta_1, \cdots, \beta_m) = \sum_{j=1}^{n} g\Big(\sum_{i=1}^{m} (y_{ij} - x_{ij}\beta_i)^2 \Big) \quad .$$

For coordinates of β_{i_1} and β_{i_2} corresponding to different equations, we have

$$E \frac{\partial^2 L}{\partial \beta_{i_1 k_1} \partial \beta_{i_2 k_2}} = 4 \sum_{j=1}^{n} x_{i_1 j k_1} x_{i_2 j k_2} E(y_{i_1 j} - x_{i_1 j}\beta_{i_1})(y_{i_2 j} - x_{i_2 j}\beta_{i_2}) \, g''(\|u_j\|^2) \quad .$$

This equals zero since the expectation equals zero conditional on $\|u\|^2$. For coordinates of β_i in the same equation, we have

$$\mathrm{Cov}\Big(\frac{\partial L}{\partial \beta_{ik_1}}\Big)\Big(\frac{\partial L}{\partial \beta_{ik_2}}\Big) = 4 \sum_{j=1}^{n} x_{ijk_1} x_{ijk_2} E(y_{ij} - x_{ij}\beta_j)^2 (g'(\|u_j\|^2))^2 \quad .$$

This has the appropriate form (3.3). Since each coordinate of u_j has the same marginal distribution, the expectation above is 4 times

$$E \, u_{1j}^2 (g'(\|u_j\|^2))^2 = \frac{1}{m} E \, \|u_j\|^2 (g'(\|u_j\|^2))^2 \quad ,$$

and the result in the spherically symmetric case follows taking inverses. For general Λ, simply transform to symmetry by $\Lambda^{-1/2}$. □

If the elliptically contoured case may be assumed, then transforming by $\Lambda^{-1/2}$ and obtaining the solution to (2.2) provides a reasonable alternative estimator depending on a one-dimensional kernel. The asymptotic covariance provided by the following result will, of course, exceed the optimal covariance above, but will generally be an improvement over applying (2.2) to the original observations.

THEOREM 3.2. *Consider the elliptically contoured case above and transform the problem so that the succinct form of model (1.1) becomes* $\tilde{y} = \tilde{X}\beta + v$ *where*

$$\tilde{y} = \Lambda^{-1/2} \otimes I \; y \quad , \quad \tilde{X} = \Lambda^{-1/2} \otimes I \; X \quad , \quad \text{and} \quad v = \Lambda^{-1/2} \otimes I \; u \quad .$$

Assume conditions A1 and A2 hold for the transformed problem. Assume also that the function ψ *is antisymmetric. Then the solution to (2.2) with y and X replaced by* $\tilde{y}$ and $\tilde{X}$ *has asymptotic covariance matrix*

$$\tilde{V}^* = \tilde{c} \; (X' (\Lambda^{-1} \otimes I) \; X)^{-1} \tag{3.4}$$

where $\tilde{c} = E\psi^2(v_{1j}) / (E\psi'(v_{1j}))^2$.

Proof. It suffices to compute the matrices Φ and R given in Theorem 2.1 for the spherically symmetric random vector $v \in \mathbf{R}^m$. By spherical symmetry, for $i \neq j$, the coordinates $(-v_i, v_j)$ have the same distribution as (v_i, v_j). Hence, for $i \neq j$,

$$R_{ij} = E\psi(v_i)\psi(v_j) = E\psi(-v_i)\psi(v_j) = -E\psi(v_i)\psi(v_j) \quad .$$

Whence, $R_{ij} = 0$. Also the coordinates of v have the same marginal distribution. Hence,

$$R(v) = (E\psi^2(v_1)) \; I \quad \text{and} \quad \Phi(v) = (E\psi'(v_1)) \; I \quad .$$

The result follows immediately from Theorem 2.1. □

Example 3.1. Consider the case of l_1 estimation. As noted in the introduction the unweighted l_1-estimator which minimizes

$$R_0(b) = \sum\sum |y_{ij} - x_{ij}b_i|$$

has asymptotic covariance matrix of the form

$$V = (X'X)^{-1}X'(\Sigma \otimes I)X(X'X)^{-1} \quad \text{with} \quad \Sigma = (\rho_{ij}\phi_i^{-1}\phi_j^{-1})$$

where $\rho_{ij} = E\,\mathrm{sgn}\,(u_{i1})\,\mathrm{sgn}\,(u_{j1})$ and $\phi_i = 2f_i(0))$. The "efficient" l_1 estimator under these circumstances solves

$$X'\Psi(R^{-1} \otimes I)\psi(b) = 0$$

with $\Psi = \Phi \otimes I$, $\Phi = (\phi_i)$, $R = (\rho_{ij})$ and $\psi(b) = \mathrm{sgn}(y_{ij} - x_{ij}b_i)$. This weighted l_1-estimator achieves the asymptotically "smaller" covariance matrix,

$$\tilde{V} = (X'(\Sigma^{-1} \otimes I)X)^{-1} \tag{3.5}$$

with $\Sigma = \Phi^{-1}R\,\Phi^{-1} = (\rho_{ij}\phi_i^{-1}\phi_j^{-1})$ as above.

For the case of the "efficient" l_1 estimator applied to the transformed problem (Theorem 3.2), we have $E\psi^2(v) = 1$ and $E\psi'(v) = 2f_v(0)$, where f_v is the marginal density of a coordinate of the transformed variable $v = \Lambda^{-1/2}u$. Hence,

$$\tilde{c} = \frac{1}{4\,(f_v(0))^2} \quad . \tag{3.6}$$

We will now compare V^* (3.3), $\tilde{V}^*$ (3.4) and $\tilde{V}$ (3.5) in the special case where $v = \Lambda^{-1/2}u$ has a multivariate t-distribution with q degrees of freedom and covariance matrix I. That is, let $u \sim \Lambda^{1/2}\,z\,/(\chi^2(q)/(m-2))^{1/2}$ where z has a unit normal distribution in $\mathbf{R}^m$. Note that while $\tilde{V}^*$ and V^* are proportional to $(X'(\Lambda^{-1}\otimes I)\,X)^{-1}$, $\tilde{V}$ has a different form. The constants of proportionality for $\tilde{V}^*$ and V^* are derived in the following proposition and efficiencies are plotted in Figure 3.1 for $q > 0$ and $m = 2, 5, 10$. For comparison, the asymptotic covariance for the least squares estimator has the same form with constant $c_2 = 1$ (the marginal variance for the transformed multivariate t-distribution), from which efficiencies of the least squares estimators were also computed and plotted. Note that although the efficiency of the l_1-estimator can tend to zero, it does so only for extreme error distributions where the asymptotic variance is already quite small.

Proposition 3.1: *For the multivariate t-distribution in m dimensions with q degrees of freedom and covariance* Λ , c^* *(3.3)* and $\tilde{c}$ *(3.6) are given by*

$$c^* = \frac{(m+q+2)(q-2)}{q(m+q)} \quad \text{and} \quad \tilde{c} = \frac{\pi(q-2)\Gamma^2(q/2)}{4\,\Gamma^2((q+1)/2)} \quad . \tag{3.7}$$

Proof. First consider the optimal covariance. Let $w = \|v\|^2/(q-2)$. Then the density of w is

$$f(w) = c(m,q)\,(1+w)^{-(m+q)/2} \quad \text{where} \quad c(m,q) = \frac{\Gamma((m+q)/2)}{\Gamma(m/2)\Gamma(q/2)}$$

and c^* will be $(q-2)$ times the value computed using this density. So the logarithmic derivative becomes

$$g'(w) = \frac{(m+q)}{2}\,\frac{1}{(1+w)} \quad .$$

Therefore, from (3.3),

$$\frac{1}{c^*(w)} = \frac{4}{m}\,\frac{(m+q)^2}{4}\int_0^\infty \frac{v}{(1+v)^2}\,c(m,q)\,\frac{v^{(m-1)/2}}{(1+v)^{(q+m)/2}}\,dv$$

$$= \frac{(m+q)^2}{m}\,\frac{c(m,q)}{c(m+2,q+2)} = \frac{(m+q)q}{(m+q+2)} \quad ,$$

from which (3.7) follows for c^* .

The result for $\tilde{c}$ follows easily from the observation that f_v (3.6) is just the density of a univariate t_q–distribution times $(q-2)/q$. □

Lastly, we calculate $\tilde{V}$ in a special case of a bivariate t-distribution. In particular, let $u_j \in \mathbf{R}^m$ be an observation from a bivariate t-distribution with q degrees of freedom and covariance matrix Λ given by

$$\Lambda = \begin{bmatrix} 1 & 2\rho \\ 2\rho & 4 \end{bmatrix} \tag{3.8}$$

for $|\rho| \le 1$. That is, let $u_j = z\,/\,(\chi^2(q)/(m-2))^{1/2}$ where $z \sim \mathbf{N}_2(0, \Lambda)$.

Proposition 3.2: *Under the above multivariate t-distribution, the asymptotic covariance (3.5) of the "efficient"* l_1*–estimator applied to the untransformed data is*

$$\tilde{V} = \tilde{c}\,(X'(\mathfrak{L}^{-1}(u)\otimes I)X)^{-1}\,, \quad \text{where} \quad \mathfrak{L}(u) = \begin{bmatrix} 1 & \frac{4}{\pi}\sin^{-1}\rho \\ \frac{4}{\pi}\sin^{-1}\rho & 4 \end{bmatrix}. \qquad (3.9)$$

Proof. We only need to compute $R(u)$ and $\Phi(u)$ as given after (3.5) for $\psi(u) = \mathrm{sgn}(u)$. Clearly, the diagonal entries of $R(u)$ are unity, and the off-diagonal entry is

$$R_{12}(u) = E\,\mathrm{sgn}(u_1)\mathrm{sgn}(u_2) = E\,\mathrm{sgn}(z_1)\mathrm{sgn}(z_2) = \frac{2\sin^{-1}\rho}{\pi}$$

from (3.8) where formula 26.3.19 from Abramowitz and Stegun (1964) has been applied. Also, since the marginal distribution of u_{1j} is the same t-distribution as the marginal for v_{1j} above, and $u_{2j} \sim 2\,v_{2j}$, we have

$$\Phi(u) = \tilde{c}^{-1/2} \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix},$$

where $\tilde{c}$ is exactly the same as in the expression for $\tilde{V}^*$. Therefore, $\tilde{V}$ has the desired form with $\mathfrak{L}(u) = \Phi^{-1}(u)\,R(u)\,\Phi^{-1}(u)$, from which (3.9) follows by direct calculation. □

Since $\tilde{V}^*$ and $\tilde{V}$ are the same except for $\mathfrak{L}$ and Λ which have the same main diagonal terms, $\tilde{V}^* < \tilde{V}$ (in the sense of having a positive definite difference) if and only if $\det(\mathfrak{L}) < \det(\Lambda)$. In fact, the ratio of these determinants is just the ratio of generalized variances, $\det(\tilde{V}^*)/\det(\tilde{V})$. Thus, $e \equiv \{\det(\Lambda)/\det(\mathfrak{L})\}^{1/2}$ is a measure of efficiency which is scaled as a ratio of variances. Direct computation shows that e monotonically decreases to zero as $|\rho| \to 1$. Furthermore, e is moderately large unless there is substantial correlation among the equations, in which case the actual variance $\det(\tilde{V})$ is small. In particular, $e \geq .82$ for $|\rho| \leq .7$ and $e \geq .62$ for $|\rho| \leq .9$.

4. l_1-Estimation of Multivariate Regression

An important special case of the foregoing results is the "efficient l_1" estimator described in the last section, whose asymptotic covariance is given by (3.4):

$$\tilde{V} = (X'(\Sigma^{-1} \otimes I)X)^{-1}$$

with $\Sigma = \Phi^{-1} R \Phi^{-1} = (\rho_{ij}\phi_i^{-1}\phi_j^{-1})$ as above. The question now arises, can one effectively estimate Φ and R in this case? We will briefly suggest some possible approaches to this problem based on recent work by Welsh (1987) and Sheather and Hall (1987). The easier of the two problems is clearly the estimation of R. The analogy principle (Manski (1988)) immediately suggests replacing the unobserved errors by the l_1-residuals obtained by minimizing $R_0(b)$ and to estimate R by,

$$\hat{R} = (\hat{\rho}_{ij}) = ((n - p_i - p_j)^{-1} \sum_{k=1}^{n} \operatorname{sgn}(\hat{u}_{ik}) \operatorname{sgn}(\hat{u}_{jk})$$

Note that we have made an adjustment for "degrees of freedom" here $(n\text{-}p_i\text{-}p_j)$ which seems natural in the l_1 case since p_i of the $\hat{u}_{ik}$'s are zero and p_j of the $\hat{u}_{jk}$'s are zero due to the l_1-fitting procedure.

It remains to estimate the elements of Φ^{-1}. The problem of estimating the reciprocal of a density function, or *sparsity* function in the terminology of Tukey, has been recently studied by Welsh (1987) and Sheather and Hall (1987). A starting point for discussion is the observation that the sparsity function, $s(u) = f(F^{-1}(u))$, is simply the derivative of a quantile function. Differentiating the identity $F(F^{-1}(u)) = u$, we obtain

$$f(F^{-1}(u))\frac{d}{du} F^{-1}(u) = 1 .$$

Thus, a natural estimator of $s(u) = \frac{d}{du} F^{-1}(u) = (f(F^{-1}(u)))^{-1}$ is Siddiqui's (1960) difference quotient of the empirical quantile function,

$$s_n(u, \delta_n) = \frac{n}{2\delta_n} \left\{ X_{([nu + \delta_n] + 1)} - X_{([nu - \delta_n] + 1)} \right\}$$

where $[x]$ is the largest integer contained in x, and $X_{(i)}$ is the i^{th} order statistic

from the sample $\{X_1, \cdots, X_n\}$. The asymptotically optimal choice of the sequence $\{\delta_n\}$ is

$$\delta_n^* = (9n^4 s^2(\tfrac{1}{2})/2(s''(\tfrac{1}{2}))^2)^{1/5}$$

see Bofinger (1975), and Sheather and Maritz (1983). In a fascinating recent paper Hall and Sheather (1987) study the distribution of a studentized quantile in which the studentization involves the Siddiqui estimator of the standard error of the quantile. They find using Edgeworth expansions that the optimal δ_n-rate to minimize size-discrepancy in associated testing procedures leads to choosing $\delta_n \sim n^{2/3}$ rather than $\delta_n \sim n^{4/5}$. These results have some support from monte-carlo investigations by Hall and Sheather (1987) and others.

Welsh (1987) has also recently investigated methods of estimating the sparsity function motivated by the problem of estimating the precision of the median and of the l_1-regression estimator. Welsh's methods may be viewed as kernel smoothing of the empirical quantile function, and thus may be expressed as weighted averages of several Siddiqui estimators. This approach seems quite promising based on the asymptotic theory. However, further experience is required to tune the choice of bandwidth and choice of kernel.

Welsh considers estimators of the form,

$$\tilde{s}_n(u,\delta_n) = \delta_n^{-2} \sum X_{(j)} h(\delta_n^{-1}(u - j/n))$$

where $h(\,)$ is the derivative of some conventional kernel, e.g. the Epanechnikov kernel,

$$h(x) = \begin{cases} -3/2x & x \in [-1,1] \\ 0 & \textit{otherwise} \end{cases} .$$

The estimator $\tilde{s}_n$ may thus be viewed as the derivative of a smoothed estimate of the quantile function. In the case of the Epanechnikov kernel we may express $\tilde{s}_n$ as a weighted average of several $\hat{s}_n(u,i)$'s,

$$\tilde{s}_n(u,\delta_n) = (3/2)\delta_n^{-3} \sum_{j \in J_n} X_{(j)}(j/n - u))$$

where $J_n = \{j \mid n(u-\delta_n) < j < n(u+\delta_n)\}$, or

$$\tilde{s}_n(u,\delta_n) = (3/n^2)\delta_n^{-3} \sum_{i}^{[n\delta_n]} i^2 \hat{s}_n(u,i).$$

Expressing the weights $w_i = 3i^2/(\delta_n^3 n^2)$ in terms of quantiles $\alpha = i/n$ we have, the weight function

$$w_n(\alpha) = 3\alpha^2\delta_0^{-3}n^{3/5} \qquad \alpha\in[0,\delta_0 n^{-1/5}],$$

where we have chosen $\delta_n = \delta_0 n^{-1/5}$, the asymptotically optimal sequence for normal errors. One may verify directly that the weight function integrates to one. In Figure 4.1 we illustrate these weight functions for several values of n. The corresponding optimal δ_n^* for the Siddiqui estimator are superimposed on this figure. In each case we have used the asymptotically optimal constants for $u = ½$, the median, and F Gaussian. It is clear that the weights implicitly assigned by the Epanechnikov kernel place most of the mass beyond the optimal Siddiqui estimator. To gauge the significance of this it is useful to consider the bias function of $\hat{s}$. Clearly, $\hat{s}(u,\delta)$ may be regarded as an estimate of the difference quotient,

$$\phi(u,h) = \frac{F^{-1}(u+h)-F^{-1}(u-h)}{2h}$$

Evaluating at the standard normal, we illustrate $\phi(½,h)$ in Figure 4.2. Comparing Figures 4.1 and 4.2 a significant bias is suggested. In Table 4.1 we compare, again for the normal case, the finite-sample bias effect,

$$s_n = \int_0^{½} w_n(h)\phi(½,h)\,dh$$

with $\phi(½,\delta_n^*)$, the bias of the asymptotically optimal Siddiqui estimator.

TABLE 4.1

Bias in Kernel vs. Siddiqui Estimators of Normal Sparsity at the Median

n	s_n	$\phi(u,\delta_n^*)$	δ_n^*/n
50	3.34	2.80	.296
100	2.95	2.71	.258
1000	2.64	2.57	.162
10000	2.55	2.53	.102
1000000	2.52	2.51	.065
∞	2.51	2.51	.000

For modest sample sizes both methods clearly over-estimate $s(\frac{1}{2})$. This is confirmed in some rather limited monte-carlo investigations. However, the kernel method is substantially worse than the simple Siddiqui estimator. Sheather and Hall (1987) argue that δ_n^* should be of order $n^{2/3}$ rather than $n^{4/5}$ if the goal is to minimize the level error for two-sided tests based on standardized quantiles, but offer no guidance on the choice of constants.

It may also be helpful to experiment with other choices of kernels. For example the biweight kernel, yields,

$$\tilde{s}_n(u,\delta_n) = \delta_n^{-2}\sum(h(i/n\,\delta_n)(2i/n)\hat{s}_n(u,i)$$

where $h(x) = -(15/16)(1-x^2)4xI(|x| < 1)$. This yields weight functions,

$$w_n(\alpha) = \delta_n^{-3}(30/4)\alpha^2(1-(\alpha/\delta_n)^2) \quad .$$

Unfortunately, they are qualitatively similar to the (more extreme) Epanechnikov case, and the finite sample biases yield similar results to those reported for the Epanechnikov kernel in Table 4.1.

5. On Affine Equivariance

To conclude, a brief *apologia* is required for the dereliction of affine equivariance. Most of the recent work on robust multivariate analysis, (see Rousseeuw(1987) and Hampel, *et al* (1986, Chapter 5) and references cited there) has restricted attention to estimators which commute with affine transformations. Suppose $T(x_1, \cdots, x_n)$ is an estimator of multivariate location based on observations $\{x_i \in \mathbf{R}^p : i = 1,...,n\}$. Then T is said to be affine equivariant if and only if

$$T(x_1A + b, \cdots, x_nA + b) = T(x_1, \cdots, x_n)A + b \qquad (5.1)$$

for any $b \in \mathbf{R}^p$ and nonsingular $(p \times p)$ matrix A. This property is particularly compelling in physical applications where, for example, the coordinate system for $\mathbf{R}^3$ is arbitrary. However, in many applications the measured coordinates *are* meaningful -- commodity bundles in economics, for example. Then, non-diagonal transformations A are difficult to interpret.

The methods suggested above satisfy (5.1) for diagonal A, and therefore *are* affine equivariant coordinate-by-coordinate. They do not commute, however, with arbitrary non-singular matrices A. Whether this failure is a mere peccadillo or a mortal sin seems debatable. Unless linear combinations of individual coordinates are meaningful quantities there appears to be little harm in restricting affine equivariance to be a coordinate-by-coordinate property. Unfortunately, the most appealing of the affine equivariant methods, due to Oja(1983) and Rousseeuw(1987) are extremely difficult to compute; this may offer another, at least temporary, rationale for the methods suggested above.

REFERENCES

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C.

Bickel, P.J. (1975). One-Step Huber estimates in the Linear Model, *Journal of the American Statistical Association,* 70, 428-433.

Blomqvist, N. (1950). On a measure of dependence between two random variables, *Ann. Math. Statist., 21*, 593-600.

Bofinger, E. (1975). Estimation of a density function using order statistics, *Australian Journal of Statistics,* 17, 1-7.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* Wiley: New York

Devlin, S.J., Gnanadesikan, R., and Kettering, J.R. (1975). Robust estimation and outlier detection with correlation coefficients, *Biometrika, 62*, 531-545.

Ibragimov and Has'minskii (1981). *Statistical Estimation: Asymptotic Theory* (tr. S. Kotz), Springer Verlag, New York.

Koenker R. and Bassett, G. (1978). Regression Quantiles, *Econometrica,* 46, 33-50.

Kuester, K. (1987) Asymptotic Consistency and Normality of Least Absolute Deviations Applied to Seemingly Unrelated Regression Systems, Technical Report: Board of Governors of the Federal Reserve System.

Manski, C. (1988). *Analog Estimation Methods in Econometrics* Chapman-Hall: New York.

Rousseeuw, P.J. (1987). Identification of multivariate outliers and leverage points by means of robust covariance matrices, Report 87-15, Faculty of Mathematics and Informatics, Delft University of Technology.

Oja, H. (1983). Descriptive Statistics for Multivariate Distributions, *Statistics and Probability Letters,* 1, 327-333.

Portnoy, S. (1977). Robust estimation in dependent situations, *Ann. Statist., 5,* 22-43.

Portnoy, S. (1979). Further remarks on robust estimation in dependent situations, *Ann. Statist., 7,* 224-231.

Ruppert, D. and Carroll, R. (1980). Trimmed Least Squares Estimation in the linear model, *Journal of the American Statistical Association,* 75, 828-838.

Sheather, S. and Hall, P. (1987). On the distribution of a Studentized Quantile, Technical Report: Australian National University.

Siddiqui, M.M. (1960). Distribution of quantiles in samples from a bivariate population, *J. Res. Nat. Bur. Standards Sect B.* 64, 145-150.

Welsh, A.H. (1987). Kernel estimates of the sparsity function, in Y. Dodge (ed) *Statistical Analysis Based on the L1 Norm,* North Holland: New York.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association,* 57, 348-368.

Zellner, A. (1962). Estimators for seemingly unrelated regression equations: some exact finite sample results, *Journal of the American Statistical Association,* 58, 977-992.

Fig. 3.1: Efficiencies for L1 and LS

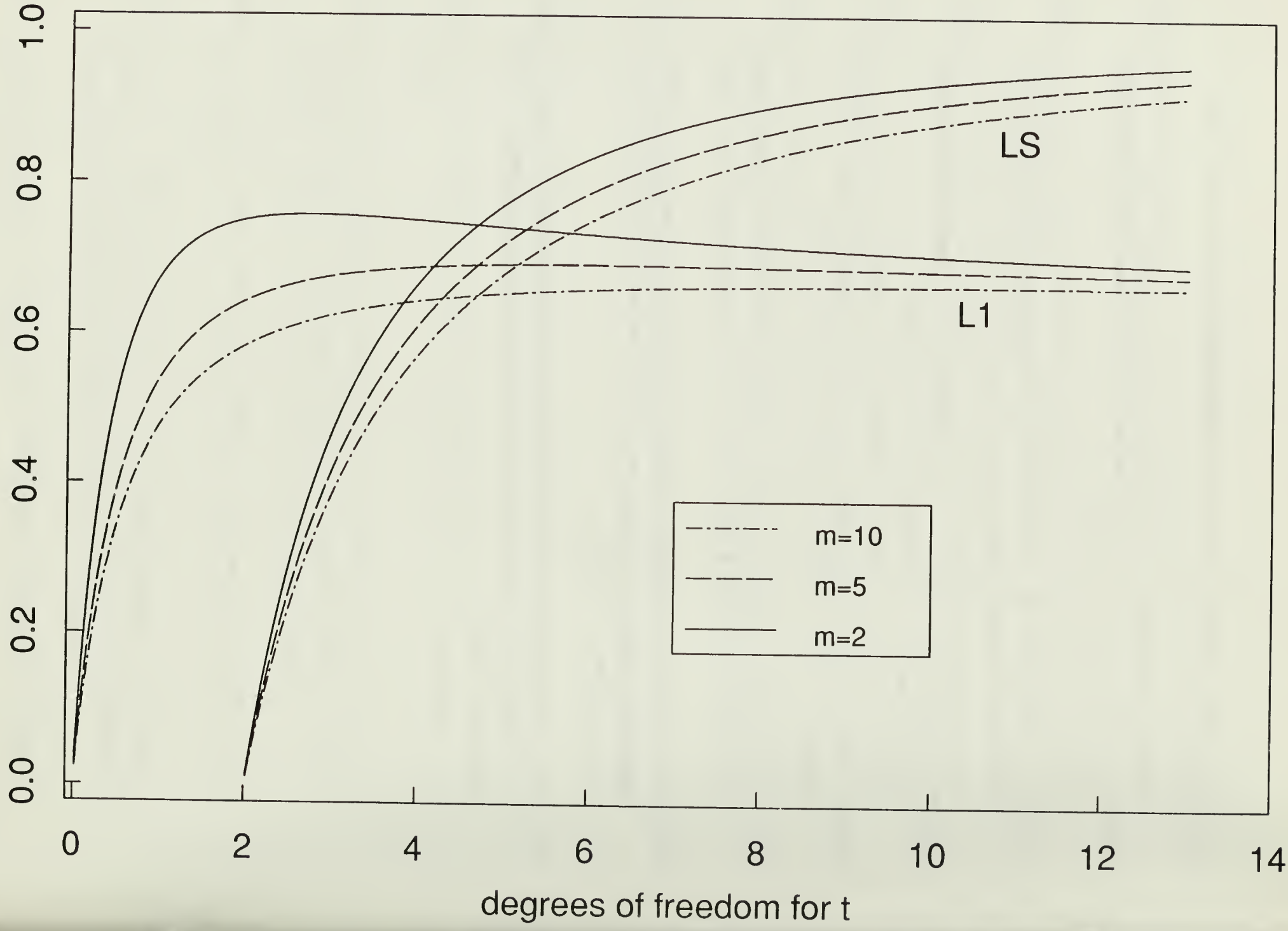


Fig. 4.1: Weight Function for Epanechnikov Kernel

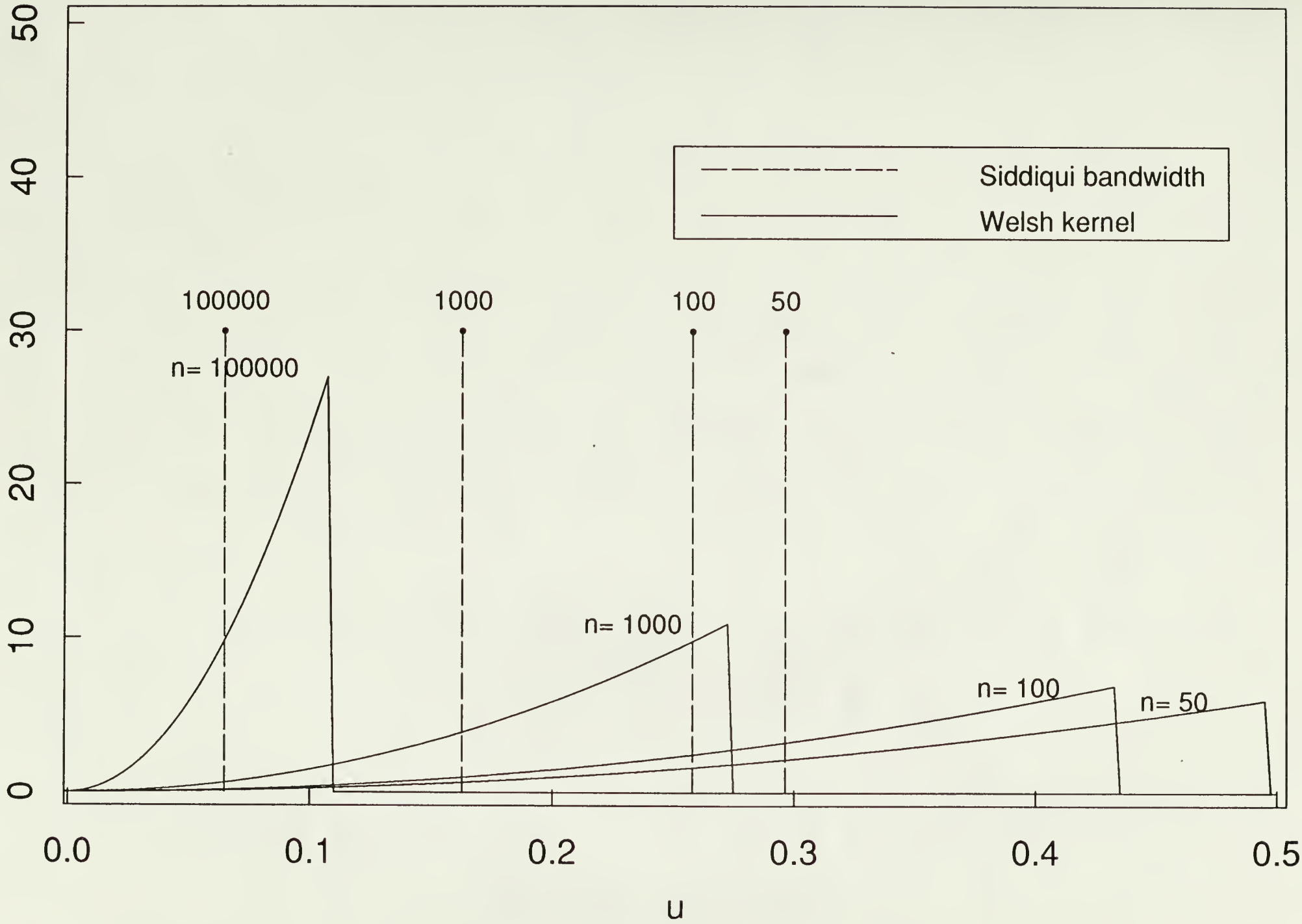
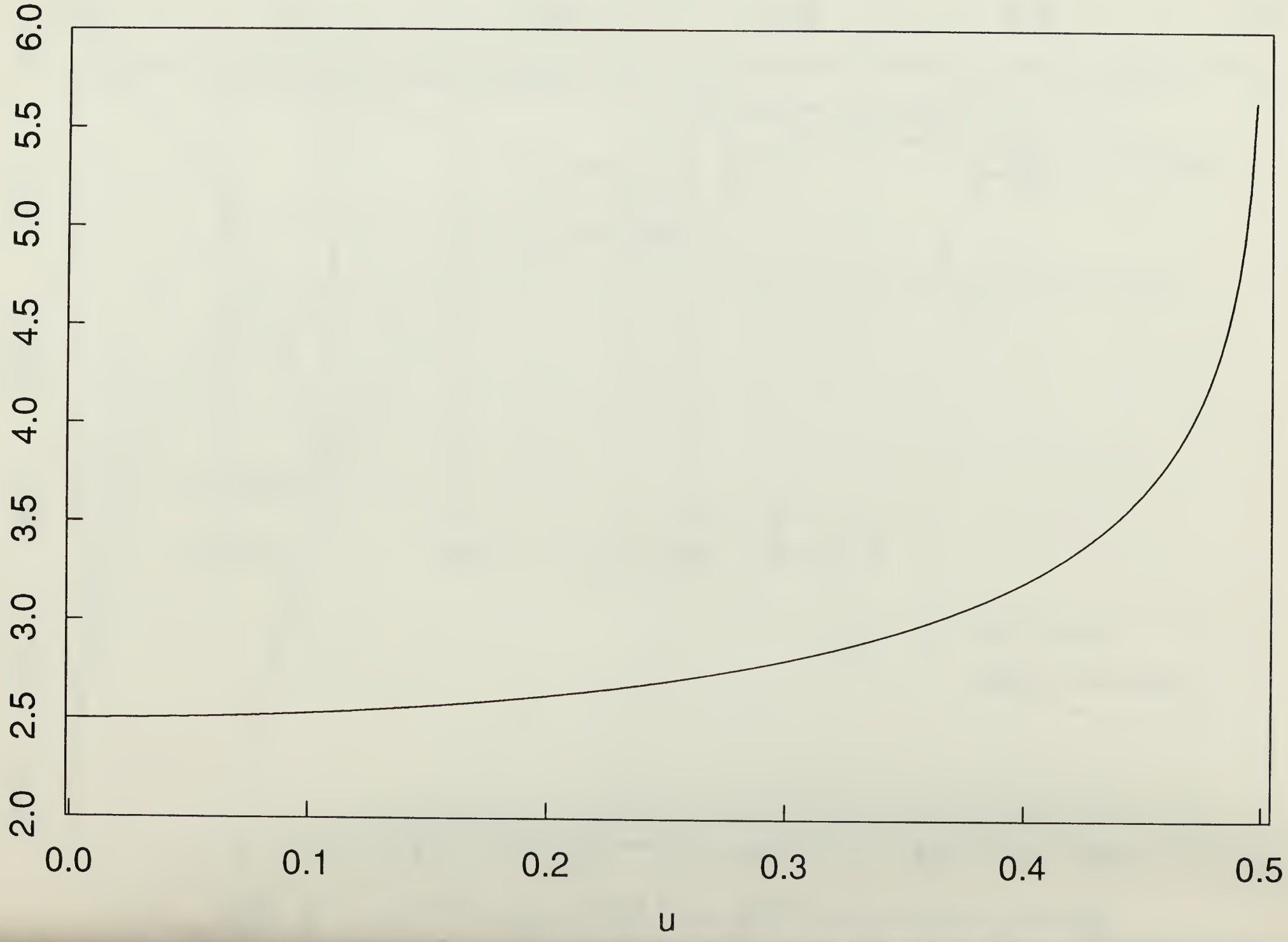


Fig. 4.2: Siddiqui Bias



UNIVERSITY OF ILLINOIS-URBANA
3 0112 060710156